Our Ref.: 2380-611

U.S. PROVISIONAL PATENT APPLICATION

Inventors: Lars Herbert BERGENLID
Magnus OLSSON

Invention: IMS VOICE SERVICES

NIXON & VANDERHYE P.C.
ATTORNEYS AT LAW
1100 NORTH GLEBE ROAD
8TH FLOOR
ARLINGTON, VIRGINIA 22201-4714
(703) 816-4000
Facsimile (703) 816-4100

SPECIFICATION

IMS Voice Services

Contents

Introduction

1.1 Purpose

In recent studies Ericsson have been investigating the technical and commercial feasibility of an early launch of non-conversational IMS solutions. Within these studies, the main guideline was that the non-conversational IMS service capabilities are complemented by conversational service capabilities in the CS domain. For many service scenarios (that do e.g., not require any tight synchronisation between the conversational and the non-conversational media) that approach turned out to be feasible.

However, the question also arised what would be the added-value to provide support for conversational, voice, services as part of the IMS at an early stage. What are the benefits and/or opportunities? How efficient would the actual solutions be? What would be the impact on later standardisation and product releases? If not feasible, which parts are still missing (e.g., as preparation for R6)

Ericsson need a good understanding of these issues as input to customer discussions, product development and standardisation.

This paper is therefore an analysis of some of the technical issues to support 'early conversational IMS' services in order to help consolidate the view on the corresponding opportunities and challenges.

1.2 Scope

The paper mainly addresses the technical aspects, focussing on the main issues to efficiently support conversational IMS services on the (radio) access side. The business and more general aspects are outside of the scope of this paper and are to be addressed in other, complementary, studies.

From a service perspective the focus shall be on voice. Most conversational IMS services would at least contain a voice element. Also from a network dimensioning perspective, the voice service would have the biggest impact near-term and will therefore to a large extend determine the feasibility of any 'early conversational IMS' deployments.

1.3 Background

It has become apparent that it will be challenging to define the full IMS functionality within the scope of 3GPP R5. A key aspect is that it is of the highest importance to ensure a high-quality standard for IMS. Since Ericsson strongly believe that IMS is a crucial element in future multimedia-enabled communication networks, it is not acceptable to rush the specifications at this stage. There is a too high risk that an immature standard would undermine the technical and commercial opportunities IMS will provide in the long run.

Based on the assumption that the complete IMS functionality needs to be distributed over at least one more release, there are discussions ongoing, both in direct customer contacts and standardisation, how this phasing shall be done. The main, near-term, IMS market drivers identified within these customer contacts are:

- support for presence and instant messaging based services to enhance the existing service portfolio

- support for IMS based (conversational) services to provide a more flexible and richer communication experience that will make it easier for operators to differentiate themselves

- support for IMS based voice services as an alternative to CS telephony services to streamline packet-switched technology based network architecture and operations

Ericsson would also like to promote early deployment of IMS solutions to allow the industry to get early experience with IMS technology, both from a technology and from an user-experience perspective.

This paper is clarifies some of the technical issues related to the feasibility to deploy conversational, voice, services based on the 3GPP R5 specifications as a complement to earlier studies into non-conversational services.

2 Concepts

2.1 Telephony

Telephony is of course the well known service that today accounts for an overwhelming part of the traffic in mobile networks. Some of the characteristics of the telephony service are:

- An extensive set of standardised services and features, some of which are strictly required by the regulators as a prerequisite for offering the service to the public.
- Fairly high speech quality both in terms of fidelity and low delay.
- Fairly high quality also in terms of response times for connection set up, service interaction etc.
- Very high efficiency in terms of radio spectrum usage and coverage.
- Wide area availability, consistent service available wherever the user is roaming

2.2 Multimedia Communication

Multimedia in this context is human to human communication established via 3GPP IMS. Voice can be, but does not have to be, one component in this communication. Some of the characteristics of multimedia voice communication are:

- Voice is only a media component, among many other. During an established multimedia session, voice communication is freely established and disconnected independently from the session itself, just like other media.
- Very little standardised services and features both when it comes to session establishment and the media itself.
- To what extent regulatory requirements will be applied to multimedia or the voice component is unclear. It is however clear that the requirements that are put on telephony cannot be applied as such.

2.3 IMS Voice Service

It would in principle be possible to use multimedia capabilities to emulate telephony like services. This would enable operators to offer a telephony like service using the PS domain and the IMS infrastructure. There are, however, a number of aspects to consider:

- Spectrum efficiency should be comparable to the CS domain service unless significant additional value can be provided by the PS domain telephony service
- Legal requirements on the telephony service must be fulfilled unless they can be fulfilled by other means
- The quality (in all aspect including service availability) of the service should be comparable to the CS domain service unless it can be offered at a lower price or together with some significant additional value
- The service level should match the CS domain service level unless it can be offered at a lower price.

3 Packet Voice Service Scenario

In this section an IMS voice service carried on the UMTS PS domain is outlined. The scenario below describes a particular profile of the IMS voice service which in this document is called packet voice service. The description will be used as a base for the analysis of a realisation and major characteristics of such a service. The description is focused on the access network aspects and simplified to the case in which the caller and called are using the same service.

The Packet Voice service is intended to a large extent replace the CS domain telephony, characterised by:

- The packet voice service is the dominant service, hence driving the network design
- It is assumed that support for the traditional CS domain telephony is still offered and is used to fulfil legal requirements, such as Emergency calls, and support for roaming subscribers.
- The service should be perceived by the end- user to be as good as, or comparable, with the traditional mobile telephony service in terms of performance and cost.

In this scenario, the main support provided by the network is:

- transport of the application flows, packets, between terminals;
- routing of the application session control packets.

As can be seen in figure 1, there are three distinctly different application flows to be considered:

- the media- speech samples carried by the RTP- protocol over UDP;
- media control- RTCP messages over UDP, and
- application control- SIP messages carrying SDP information over UDP.

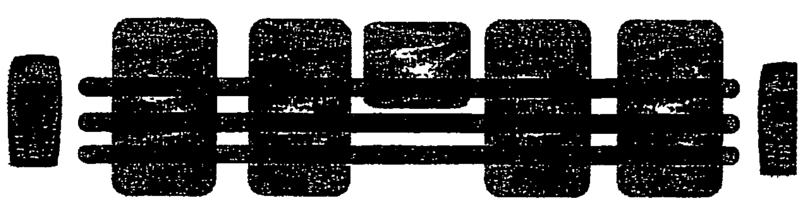


Figure 1 Basic Packet Voice Application Flows

The overall requirement from the end- user on the performance and cost of the packet voice service is reflected by requirements on how the network route and transport each of the individual flows.

Each of the flows affect the overall service behaviour, performance. For example, the media flow is significant in when it comes to providing a good speech quality and the application control flow has a significance when it comes to e.g., service set-up time.

In addition, the requirements of the operator on a reasonable provisioning cost also comes into the picture. In difference to over- provisioning access networks, such as LAN's, cellular is about providing seamless, wide- area coverage at a reasonable cost. In contrast to over provisioning access technologies, one cannot 'throw bandwidth' at the problem, since this will lead to unreasonable provisioning costs.

4 Overall Architecture and realisation

4.1 Overview of the System Realisation

When using the WCDMA technology to access the SIP based IP multimedia system, the SIP signalling and the media streams must be transported over the Universal Terrestrial Radio Access Network (UTRAN) and Packet Switched domain (PS domain). Efficient usage of resource is enabled with careful choice of the configuration used to transport the IMS bearers, including the SIP signalling and media, over the Air interface.

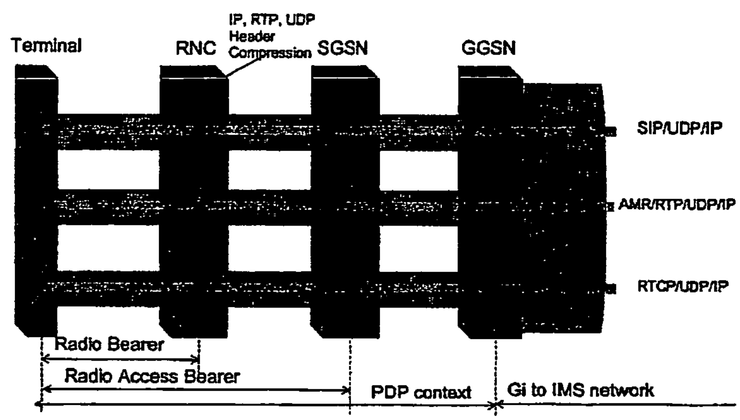


Figure 2 Example of bearers for Packet Voice

The mobile chooses the means by which the application flows IMS bearers should be transported over the PS domain. This is dependant upon the user requirement, application, the availability of radio bearers, charging considerations. Figure 3 shows a voice communication using the SIP based IP multimedia system. The SIP/SDP signalling, the AMR (RTP) media and the RTCP signalling has their own PDP contexts (and hence Radio access bearers). The PDP context for SIP/SDP signalling is always active to enable the mobile to receive IMS sessions, but during idle periods the RAB can be deactivated.

The mobile selects the configuration of the access network required to supports the communication media's negotiated through the SIP signalling. When establishing the required configuration, the mobile provides the required Traffic flow template (TFT) information to the GGSN, instructing the GGSN to place the correct flows in the selected PDP contexts. The mobile must hence know how to map the signalling and media flow to PDP contexts.

4.2 Assumptions

The basic assumptions is that the packet voice service is defined within the limits of the 3GPP Release 5 standards including referenced IETF RFCs. This means e.g., that IP header compression as well as the SIP signalling compression is supported while some other optimisations like Unequal Error Protection and Rate Control are not supported.

To further simplify the packet voice service it is assumed that only the AMR 12.2 kbps mode is used.

The packet voice service shall be compliant with 3GPP Release 5 SIP signalling including extensions for e.g., SIP compression and Security.

It should be noted that 3GPP and IETF is still working on the required SIP extensions and the messages contain variable length fields, hence the message size can not be exactly determined. In the characteristics analysis below the message sizes have estimated conservatively.

It is assumed that the packet voice media and signalling flows will require 3 separate PDP contexts and the associated RABs

- One PDP context for SIP/SDP
- one PDP context for AMR (RTP)
- one PDP context for RTCP

It would have been natural to put the RTCP flow on the same RAB/PDP context as the SIP/SDP flow or the AMR (RTP) flow, but it does not seem possible. Combination of the AMR (RTP) flow and the RTCP flow would result in excessive use of radio resources. Combination of the SIP/SDP flow and the RTCP flow would conflict the charging requirement that the GGSN shall be able to ensure that the PDP context used for signalling is only used for signalling to and from the P-CSCF.

Note that changes in the standards or operator requirements may influence the assumptions.

4.3 Overview of the information flows

SIP/SDP

SIP/SDP signalling is used for multimedia session control. Some aspects of the service behaviour depend on the QoS given to the SIP/SDP signalling, e.g., service availability and response times. The session establishment time for example is directly proportional to the delays experienced by the SIP/SDP messages.

SIP signalling is a request - response type of communication with typical packets sizes in the range 300 – 900 octets. Session set up involves a rather big number of transactions.

The number of messages that are exchanged varies, but for a straightforward session set up there are typically 11 messages with a total volume of minimum 7-8 kbyte.

SIP/SDP signalling compression can be used to reduce the SIP messages and thereby reduce the transmission delay. SIP/SDP messages are transferred compressed between the UE and the P-CSCF. Compression is therefor transparent to the PS domain and the radio networks. With compression the SIP/SDP message volume can be reduced (normally the compression would be approximately 4 times). However, the setup time is still influenced by the number of roundtrips.

IP header compression will further reduce the amount of information that needs to be transferred over the air interface.

There are three QoS aspects of special importance for the SIP/SDP signalling traffic, when used for the packet voice service:

- Bit rate : Signalling is low volume traffic with a low demand of average bandwidth. However in order to keep delays down, the transmission rates should be higher.
- Delay : Total delay for a message is the important QoS parameter. It is partly dependent on the available bit rate but also influenced by bearer handling delays and retransmissions over the air interface.
- Priority : To ensure a good service behaviour also under heavy load conditions, signalling traffic should get priority.

Of the available QoS classes, the interactive class is the one that is most suitable for this type of traffic.

SIP signalling should not be put on a UMTS bearer that carries large volume user data, like a bearer used for web browsing.. If the same bearer is used, there is no way today to prioritise SIP messages or other signalling.

Therefore a dedicated bearer should be established for SIP (and other similar signalling). In this way the signalling does not have to compete with other traffic.

But even with a dedicated bearer there may be a need for QoS improvements for the application signalling when using an interactive bearer, and this is why a Signalling Traffic bearer currently is discussed within 3GPP.

RTP flow

The RTP flow carries the media flow, which in this case is the 12.2 kbps AMR coded speech. In addition to the AMR coded speech the RTP payload also contains a payload descriptor which gives a total payload of 32 octets.

The overall IP packets for the media flow are AMR/RTP/UDP/IPv6, where the total header information is 60 octets, which then gives a total message size of 92 octets.

For the RTP flow the proposed solution is non-optimised in the sense that an unequal error protection (UEP) is not included. For the inclusion of UEP the desired approach is to use UDP lite which is not available in a 3GPP release 5 time frame.

In addition to the speech packet there are also the SID packets (7 octets payload) and DTX.

For the RTP flow a conversational RAB is the best choice, and this conversational RAB has to be capable of transferring 92 octets every 20 ms, i.e., a data rate of 36.8 kbps.

RTCP flow

The RTCP flow have messages which are usally a bit larger than the RTP messages. The RTCP flow messages are also transferred rather infrequently. To allocate a conversational RAB for this flow would result in an unnecessary capacity reservation and a better approach seems to be to use an interactive RAB for the RTCP flow. This interactive RAB would be allocated at the same time as the conversational RAB for the RTP flow, which means that this interactive RAB will be transferred on a dedicated channel, i.e., it will to some extent benefit from the performance of a dedicated channel.

4.4 Signalling flow overview

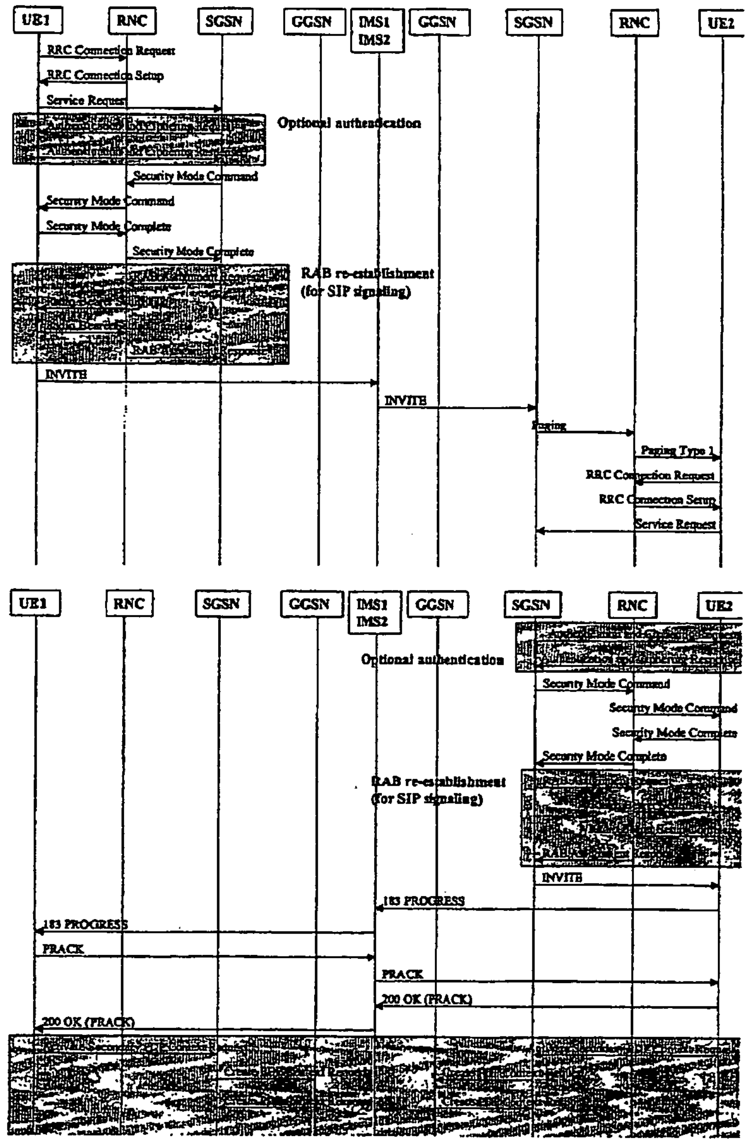
This section chapter shows one example of a session establishment that only involves a speech component.

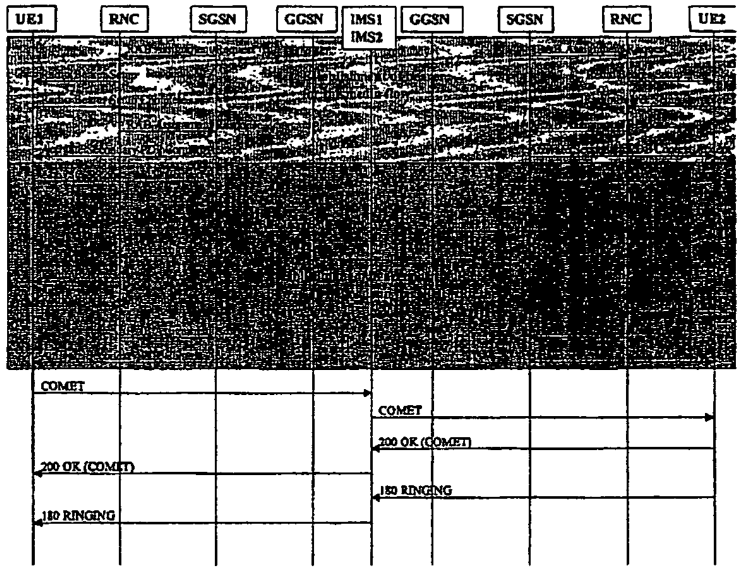
It is assumed that both UE1 and UE2 are already registered on the IMS level and that they also have PDP contexts for signalling. The RABs have though been released due to inactivity.

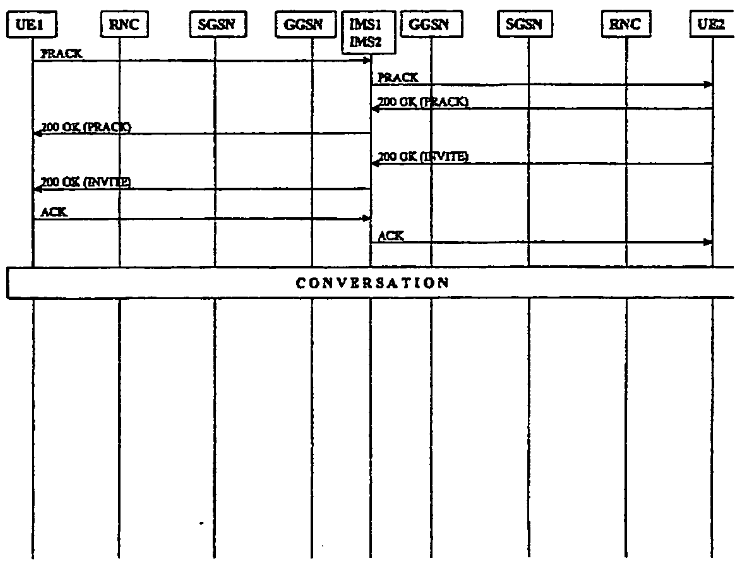
Flow outline:

- UE1 starts with a re-establishment of the RAB
- When the RAB re-establishment is finished UE 1 can send the INVITE message towards UE2
- Since UE2 does not have the RAB for the Signalling PDP context active the SGSN will page UE2 who will then re-establish the RAB.
- The UE2 can now receive the INVITE and will respond with a Progress indication
- UE1 will then send inform UE2 that it requires resources to be required for the RTP (AMR) and RTCP flows to continue.
- UE1 and UE2 activates PDP contexts for the RTP (AMR) and RTCP flow
- When UE1 has the required resources it informs the UE2 and UE2 responds when it has got the required resources.
- UE2 sends ringing to UE1 and the RTP (AMR) and RTCP flows can start.

The example flow is shown with some more details in the following figures. Note, however, that some messages has been omitted for simplicity (e.g., HSS interactions, Go signalling,...).


UE1
RNC
SGSN
GGSN
IMS1
IMS2
GGSN
SGSN
RNC
UE2
RRC Connection Request
RRC Connection Setup
Service Request
Optional authentication
Security Mode Command
Security Mode Command
Security Mode Complete
Security Mode Complete
RAB re-establishment
(for SIP signaling)
INVITE
INVITE
Paging
Paging Type 1
RRC Connection Request
RRC Connection Setup
Service Request
UE1
RNC
SGSN
GGSN
IMS1
IMS2
GGSN
SGSN
RNC
UE2
Optional authentication
Security Mode Command
Security Mode Command
Security Mode Complete
Security Mode Complete
RAB re-establishment
(for SIP signaling)
INVITE
183 PROGRESS
183 PROGRESS
PRACK
PRACK
200 OK (PRACK)
200 OK (PRACK)


UE1
RNC
SGSN
GGSN
IMS1
IMS2
GGSN
SGSN
RNC
UE2
COMET
COMET
200 OK (COMET)
200 OK (COMET)
180 RINGING
180 RINGING


UE1
RNC
SGSN
GGSN
IMS1
IMS2
GGSN
SGSN
RNC
UE2
PRACK
PRACK
200 OK (PRACK)
200 OK (PRACK)
200 OK (INVITE)
200 OK (INVITE)
ACK
ACK
CONVERSATION

5 Radio Network Realisation for the support of packet voice

5.1 Overview

The following is an overview of how a conversational IMS speech service for a 12.2 kbps AMR coded speech could be realised in a WCDMA radio access network, with the overall assumption that the solution is based on 3GPP release 5. This means that ROHC header compression is available, but not the further optimisation features that are expected in 3GPP release 6 and onwards.

For the first PDP context for the SIP signalling an interactive RAB is the natural choice. The establishment of this interactive RAB is done in the same way as in earlier releases. Also the establishment of a second interactive RAB is a rather straight forward solution.

One beneficial approach is to establish plural new PDP contexts for a session using a single PDP context request. This procedure may also be beneficial in the situation where there is only a single media flow because established a single media flow may require several PDP contexts.

Then when it comes to the establishment of the conversational RAB, there are a couple of options to choose among. In the 3GPP release 5 there will not in the RAB assignment be any detailed SDU information, instead the RNC will receive a RAB assignment with e.g max SDU size, a bit rate that corresponds to the uncompressed RTP/UDP/IP flow and with a source descriptor that indicates speech

- One of the possibilities which seems to be the most natural, is to follow what is requested in the RAB assignment. This gives the possibility to transfer the ROHC compression synchronisation information, i.e., the additional information that is needed until the header compression functionality has started to compress the RTP/UDP/IP flow, without segmentation and the corresponding build up of speech packets in the RNC. This however requires that a lower spreading factor is used initially and that a channel switching is done after a couple of frames when the header compression has synchronised.
- Another possibility is to depend on that the source descriptor indicates speech, and assign already from the beginning a higher spreading factor. In this case the longer initial frames has to be segmented which can be done either by using the unacknowledged RLC mode or by using the segmentation capabilities in the ROHC header compression. In this case there is a need for an active buffer management in the RNC, otherwise an additional speech delay can be created. This overall solution however means that it is already in the RAB assignment taken for granted that the actual compression possible without really knowing this. Ericsson think that this kind of approach should be better supported in the standard before it is implemented and that this should be part of the 3GPP release 6 and onwards.

5.2 RABs and RAB realisation

From what can be seen from the overview above and from the signalling flow, the following RABs and RAB combinations has to be supported

- Interactive RAB for the SIP signalling
- Two simultaneous interactive RABs for SIP and for RTCP
- Simultaneous interactive RABs for the SIP and a conversational RAB for the RTP flow

- Simultaneous two interactive RABs for the SIP and RTCP flows and a conversational RAB for the RTP flow

Two of these RAB combinations, i.e., the two interactive and the simultaneous interactive and conversational are only needed during a short transition time in the transition phase from the first interactive RAB for signalling to the state where all the three RABs are established.

Header compression

The ROHC header compression will compress the RTP/UDP/IP header from 60 octets to normally 3 octets. The reason for not compressing it further down to normally 1 octet is due to that it is not possible to switch off the UDP checksum in IPv6. This can most likely be improved when UDP lite is introduced.

This means that before UDP lite is available, a 12.2 kbps AMR coded speech service, will be optimised for 35 octets, but it also be ensured that a number of different sizes up to the max number of octets are possible. When more information shall be transferred, channel switching can be used in order to temporarily increase the bit rate. In order to reduce the number of formats the possibilities within ROHC to limit the number of sizes of the header may also be used. This may give the possibility to use the transparent RLC mode. Otherwise the less efficient unacknowledged RLC mode must be used instead

RAB realisation for the SIP/SDP signalling flow

For the signalling flow the already available interactive RAB is a suitable choice, i.e., an interactive RAB combined with a 3.4 kbps signalling radio bearer. With the channel switching functionality this RAB has the properties that the data rate is adapted to the amount of data that needs to be transferred. This adaptation also takes into account both the coverage aspects and also the current loading in the cell

RAB realisation for the simultaneous SIP/SDP and RTCP flow

Also for the RTCP flow the already existing interactive RAB seems to be the natural choice. For the combination of two interactive RABs our view is that a combination on the MAC level is appropriate. This gives the characteristics that the available bit rate is shared between the two flows and where one flow can use all the available data rate in case the other flow has no data to transmit. Also this RAB combination will have the same adaptation properties as is described for the single interactive RAB above, in which case the combined data rate is controlling the overall data rate.

RAB realisation for the simultaneous SIP/SDP and RTP flow

For the RTP flow a conversational RAB is needed. For the initial frames, i.e., during the initial header compression synchronisation, there is a considerable larger amount of data to be transferred than during the normal state. And the proposed way to handle this is a 3GPP release 5 time frame is to make the RAB realisations for the RTP flow with two different spreading factors, one that is optimised for the compressed RTP/UDP/IP flows in order to efficiently utilise the resources and one that is used in order to transfer the larger sizes during the synchronisation of the header compression. This however will

have a clear disadvantage from a link budget point of view. For the combination of the conversational RAB with the interactive RAB there is two separate transport channels, and then there is an additional transport channel for a 3.4 kbps signalling radio bearers in the same way as for the interactive RAB above. An important issue here is to limit the amount of data within the bounds given by the selected radio bearer, which is done by an appropriate selection of the allowed transport format combinations.

RAB realisation for the simultaneous SIP/SDP, RTCP and RTP flow

The combination of the SIP, RTCP and RTP flow will be achieved by adding an interactive RAB to the realisations of the combination of a conversational and interactive RAB described above, i.e., two RAB realisations with different spreading factors as described above will be needed.

As the requirement differs for the RTCP and SIP flows while at the same time the total bit rate should be limited so that the RAB realisations for these flows can be made with suitable radio bearers, most likely for these RAB combination will have different transport channels for the two interactive RABs. The limitation of the total bit rate is done by limiting the allowed transport format combinations.

The overall RAB realisation for the simultaneous SIP, RTCP and RTP flow will them be a combination of two interactive and one conversational RABs together with the signalling radio bearers. The overall RAB realisation will have 4 transport channels, one for the conversational RAB, one for each of the two interactive RABs and one for the signalling radio bearers.

6 Core Network Realisation for the support of packet voice

The main task of the core network in the IMS service context is to handle signalling and voice traffic efficiently across an IP network that also transport a wide variety of other data sessions with different real-time or non real-time requirements. The core network will, at the edge of the network (GGSN), ensure correct mapping of the data flows into the appropriate GTP tunnels and ensure that agreed QoS are maintained. Since the 3GPP R5 standard is still work in progress, the description below is preliminary and may be changed to adapt to the final version of the standard.

For IMS access, the terminal must establish PDP context towards the appropriate APN and be assigned (by GGSN) an IPv6 address with an address prefix related to the IMS domain. (The 3GPP R5 standard allows for stateful address allocation and stateless address allocation with one /64 Ipv6 address prefix per PDP context).

Identification of the PDP context for SIP signalling bearer is not yet defined in the standard. However it is expected that all signalling is transferred on a specific PDP context to allow the operator to select free transport or special charging for signalling as well as special policing/filtering towards IMS. GGSN will filter the traffic from this bearer and only allow traffic towards P-CSCF(s) (identified by the IPv6 address) on this PDP context. The P-CSCF IPv6 address (es) is expected configured through the APN specification for the GGSN. The signalling bearer can be of type interactive. In the core network and also on the network between GGSN and P-CSCF, the QoS on IP level may use DiffServ class "assured forwarding" with low drop precedence. The mapping will be generally configurable per APN and for the Gn network. In the uplink direction the signalling flow will be mapped into the signalling PDP context tunnel by use of the TFT

filter in the GGSN. P-CSCF IPv6 address is the natural IP header element used in the filter.

When the terminal establish sessions by using SIP signalling, equivalent secondary PDP contexts will also be established by the terminal. These PDP contexts will be correlated with the SIP sessions through signalling across the Go interface (GGSN <- >P-CSCF/PCF) (binding information exchange). The GGSN creates a GPRS related charging ID per PDP context that will be used both by the SGSN and the GGSN for CDR correlation. This ID is also exchanged across the Go interface for GPRS to SIP charging correlation. There are also additional requirements to co-ordinate the sessions on SIP level and GPRS bearer which requires the PCSCF to inform GGSN of changes on the session level and vice versa like session release and for GGSN to inform P-CSCF of loss of radio bearer etc. Currently, in the PS core network there are no such session and bearer level co-relation performed.

QoS and session destination information exchange is also expected across the Go interface and the GGSN will use this information in the admission control for the PDP context and also to filter traffic for the SIP related PDP contexts. The filtering is used to secure correct destination and source for the traffic in the SIP allocated PDP context. Mapping from external traffic to PDP context will be done using the TFT specified for the context. Policing on PDP context level will also be used to prevent accidental or malicious flooding etc. of the PDP contexts. The traffic will be marked on the IP level with DiffServ classes according the agreed QoS through PDP context set-up and Go information exchange.

In the GGSN and in the rest of the network the SIP traffic will be mixed with other type of traffic on different QoS levels. The larger amount of the data traffic is expected to be of type background (best effort) with very different real time requirement to the voice related traffic. The Core and external network nodes will use the QoS information in the IP header to forward packets according to QoS agreement. DiffServ queuing, scheduling and drop techniques are used in the forwarding. Voice traffic is expected to be marked with EF (expedited forwarding) class and will be secured low delay and low loss in the core network if the right dimensioning and service level agreements (SLA) are secured through node and network configuration.

The nodes offered for the core network are designed to be able to handle the traffic with high capacity and low delay. The specific requirements for the SIP traffic with high control message rate (PDP context combined with Go traffic) and advanced requirement on traffic filtering etc. introduces extra strain on specially the GGSN. To handle this, the GGSN and the routers are built including hardware support to handle forwarding and QoS with low node delay for the prioritised traffic.

From the discussion above, it is visible that there are specific functional requirements on the existing Packet Core Network in order to deliver IMS services.

- Core Network (SGSN & GGSN) need to be aware of the PDP context used for SIP signalling
- Minimum number of PDP contexts for every Packet Voice session is 3
- Charging correlation handling over Go interface and handling session and bearer level interaction described above
- SDP media type agreed on the session level are enforced in the PDP contexts/bearer via the media binding parameter
- In order to perform the session and bearer correlation function, the media binding parameter need to be transferred from the session level to the PS domain via the terminal
- Discovery of P-CSCF address via PS domain mechanism need to be resolved
- Policing in different level adds to the complexity of the overall system designed for general packet services

- Charging principles need to be finalised
- In addition, there are open items related to Service Based Local Policy functions such as 'open/close gate' that need to be addressed beyond Release 5 timeframe.

7 Characteristics and Possible Enhancements

The following are some general characteristics regarding capacity, coverage and setup delays for the outlined packet speech service compared with the more classical circuit switched speech service.

The overall radio capacity will be lower than for a circuit switched speech service (initial estimates show a 25% decrease). The following are some examples of areas where activities and improvements related to the capacity are expected.

- By introducing UDP lite the ROHC header compression will be improved
- UDP lite will also give the possibility to make use of a speech frame that has bit errors in the less important sub flows
- By introduction of unequal error protection the requirements of the different sub flows can be optimised individually which gives the possibility improve the overall capacity.

The coverage it will be lower than for a comparable circuit switched speech service. The most important improvements in this area is to make it possible to transfer the initial long headers without the reduced spreading factor. Examples of possibilities are

- Make a segmentation of the initial longer frames
- Reduce the initial frames by not having any payload in them

Regarding the PS Core network it can be noted that 3 PDP contexts would be needed per packet voice session and hence the network must be dimensioned to handle the PDP contexts and the associated signalling.

Regarding the call set up times, an average call set up time (up to ringing) for a mobile to mobile call is estimated to be a little bit more than 10 seconds excluding the authentication. Further work will of course be made in this area in order to improve this.

8 Summary and Conclusion

From what is described above the overall conclusion is that it is technically possible to have a radio network that supports an AMR 12.kbps packet speech service based on 3GPP release 5. However from the above:

- There are a number of assumptions that has been made, and all these assumptions has of course to be more thoroughly worked through
- There are a number of areas where further work are clearly desirable in order to improve the performances, and this is also what is currently taking place in the 3GPP standardisation
- There are also a number of different decisions that has to be made in order to define such service. These decisions has to be made on a global scale in order to secure the interoperability and the availability of terminals (e.g., for the packet voice service the application/terminal must have enough knowledge to activate separate PDP contexts for the RTP, RTCP and SIP/SDP flows).

- The basis for these decisions will change as the 3GPP standard is evolving, i.e., several of the selections of the different alternatives would be different when based on 3GPP release 6 instead of release 5.

This means that a packet voice service based on 3GPP release 5 will most likely not happen as a mass market service. Instead the main stream solutions will be based on later 3GPP releases where all the main issues has been settled.

The design of a system solution for an optimised mainstream IMS voice service will be an iterative process where all parts must be taken into consideration. IMS is one important subsystem in the service delivery. However, it must also be understood how the application, UE, UTRAN and PS CN works together with IMS to deliver a consistent service to the end user. The total system view has so far been somewhat neglected in 3GPP and hence it is very difficult to see how optimised mainstream IMS conversational services can be built based on release 5.

It should also be noted that this paper only addresses a part of the whole solution, a number of issues to account for:

- simultaneous use of packet voice and other services
- what happens when the system is coverage limited rather than capacity limited
- how to ensure packet voice operation in high load conditions

9 Way Forward

The packet voice service scenario is Ericsson's interpretation of a service that could be interesting for operators. Further Ericsson has made estimated what will be included the 3GPP release 5 standards. Any changes in the service scenario, the standards or other assumptions may very well influence the design choices and hence the implications in terms of capacity, coverage and service quality.

Ericsson would like to further investigate possible IMS service scenarios including voice and would encourage broad participation in continued discussions to aid all parties understanding of the implications on the Applications, Terminals, UTRAN and PS CN in terms of both standards and products.

SIP Mobile Multimedia System Description - PA5

1 Introduction

This paper describes on a high level the Ericsson solution of a complete mobile multimedia system for supporting different media using IP technology.

It starts with Section 2 is an description overview of the architecture for the complete mobile multimedia system. The main focus is on the IP Multimedia Subsystem (IMS) architecture. The other sub-system components are also described but the focus is then on the expected functionality needed to support IMS. The main drivers for introducing IP technology into mobile networks: new services and a common transport technology. It then gives an overview of the mobile multimedia system focusing on the core network architecture.

Section 3 describes system concepts that are necessary or adds value to IMS.

Section 4 gives a high-level view of what products Ericsson can offer for a complete mobile multimedia system.

Section 5 gives a brief migration view how an early IMS system can be migrated to real-time conversational SIP based IP multimedia system.

2 System Architecture and Interfaces

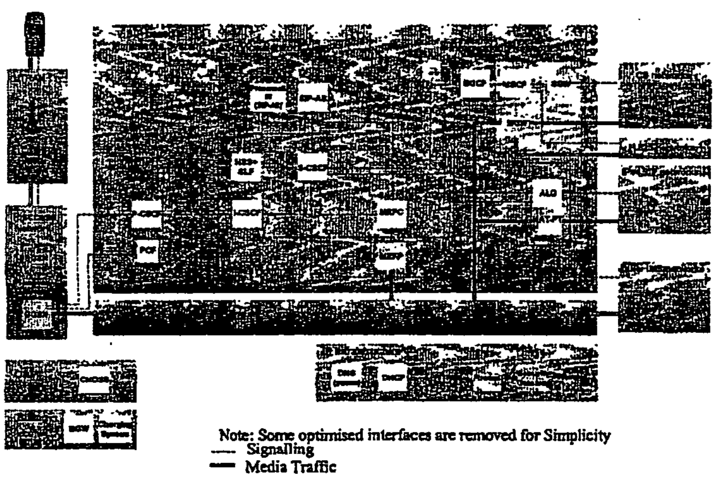


Figure 1 System Overview

Figure 1 shows a high level view of the system components required in order to deliver services to a SIP mobile multimedia system. The role of these system components is described below.

2.1 Terminals

2.1.1 Terminal Functionality

In the same manner in which the terminal will connect to different core networks, there will be a number of terminals which can also connect to the IMS network. Terminals connecting to the IMS network will have a greater role to play in supporting the end-users' communication needs and will support all or parts of the following functionality:

- SIP/SDP as defined by IETF and 3GPP.
- IPv6 profile for cellular networks for IMS
- Multiple simultaneous PDP contexts (and RABs).
- Order mapping of the media flows to the correct PDP contexts.
- Conversational QoS.
- Header compression (ROHC)
- SIP header compression
- Java Based terminal service creation environment.
- SIP based presence and instant message applications.
- IPv4 and IPv6 for non-IMS applications
- Support for non-IMS application (e.g. WAP 2.0, MMS)
- Ability for the IMS application to invoke other applications (e.g. Streaming)
- Integrated camera
- Security functions (TLS for WAP, IMS AKA for IMS).

2.1.2 Terminal types

The terminals accessing IMS are segregated into two basic types – the smartphone and the split user equipment.

The smartphone is an tightly integrated device containing the UMTS radio functionality and the IMS application. This is an evolution of the terminals that mobile users are familiar with in 2G networks. The level of functionality supported in the smartphone is dependant on the market segment the phone is aimed for, but may include applications streaming, camera, WAP and messaging. These phones are mass market phones.

The split user equipment has two components. One component is a device containing the UMTS radio functionality, and may support a number of applications, however the IMS application resides on another device. This component could be a laptop or a PDA, and may cover e.g. Microsoft XP Messenger, which is gaining popularity as a SIP based application.

2.2 UTRAN

The UTRAN plays the role of controlling the radio network resources. In order to support real-time conversational SIP based IP multimedia the base stations and RNC

are expected to provide new features. Some of the more important functionality expected are listed below:

- Header compression (ROHC)
- Conversational RAB for the transport of Voice and video media
- Interactive RAB for the transport of IMS signalling (SIP signalling, and possible other supporting signalling). This might be in the form of a Signalling RAB when/if standardised.
- Transport for RTCP signalling.
- Flexible RAB combinations allowing simultaneous flows for media, signalling and other services (e.g. WAP). The number of simultaneous flows is dependant upon the end-users service experience that will be provided to Hutchison's subscribers.

2.3 IP Access Network

The main task of the IP access network in the IMS service context is to handle signalling and voice traffic efficiently across an IP network that also transports a wide variety of other data sessions with different real-time or non real-time requirements. The core network will, at the edge of the network (GGSN), ensure correct mapping of the data flows into the appropriate GTP tunnels and ensure that agreed QoS levels are maintained. Since the 3GPP R5 standard is still work in progress, the description below is preliminary and may be changed to adapt to the final version of the standard. For IMS access, the terminal must establish a PDP context towards the appropriate APN and must be assigned (handled by GGSN) an IPv6 address.
Identification of the PDP context for SIP signalling bearer is not yet defined in the standard. However it is expected that all signalling is transferred on a separate (?) PDP context to allow the operator to select free transport or special charging for signalling as well as special policing/filtering towards IMS. GGSN will filter the traffic from this bearer and only allow traffic towards P-CSCF(s) (identified by the IPv6 address) on this PDP context. The P-CSCF's IPv6 address(es) towards the terminal is expected to be configured through the APN specification for the GGSN. The signalling bearer can be of type interactive.
When the terminal establishes sessions by using SIP signalling, it will also establish secondary PDP contexts for the transport of the bearers.

2.3.1 GGSN

The GGSN provides an interface between mobile radio core networks (GPRS and UMTS) and other packet data networks, such as the Internet, corporate intranets, and private data networks. In this role, the GGSN is responsible for session management within the mobile network, as well as for encapsulation and de-encapsulation of bearer traffic sent to and from Serving GPRS Support Nodes (SGSNs). The GGSN serves as the IP access node for IMS.

Examples of GGSN functions that will or might be used to support IP multimedia are:

- VPN (IPSec and MPLS)
- IPv6
- Routing protocols (OSPF, BGP-4, IS-IS etc)

- QoS (multiple mechanisms (e.g. diffserv) applied at the per-PDP-context level and at the aggregated-trunk level)
- Go functionality required to support charging..
- Multiple PDP contexts per subscriber. Each context will then have a separate IPv4 or IPv6 address.
- Several secondary PDP contexts per primary PDP contexts. By using a secondary PDP context a user will be able to have more than one PDP context per IP address, with a different level of QoS if necessary.
- Security (e.g. IPSec, IKE)

2.4 IP Multimedia Subsystem

Ericsson's network solution for IMS is based on the separation of functionality into a control layer and a connectivity layer as shown in Figure 2.

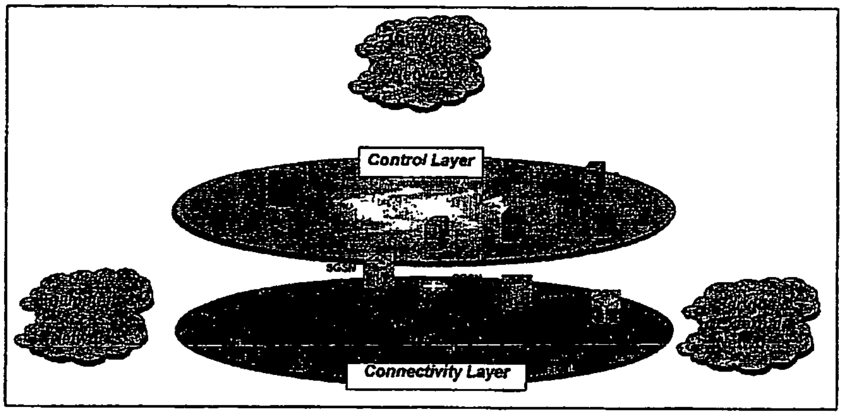


Figure 2 Ericsson high level solution for IMS

The control layer hosts control servers such as the IMS servers (CSCF, MGCF, BGCF, SGW and MRFC). The control servers may also handle functions like mobility management, security, charging, conferencing and interworking towards external networks on control plane level. The HSS is the subscriber master database for the PS domain and IMS. The main protocol in the control layer is SIP. SIP allows several transport protocols, UDP is mandatory, which is supported by the Ericsson products.

The connectivity layer is used for media and includes media gateways and media mixing functions. The media gateways provide different kinds of interworking for the media, including interworking between different transmission technologies and interworking between different media formats.

The interface between the control layer and the connectivity layer consists of plain IP routing or gateway control protocols such as H.248. MRFC and MGCF use gateway control protocols to control media manipulation and interworking in MRFP and MGW.

The IP access gateway (GGSN) is considered to be part of the connectivity layer. Although it contains some control functionality, the IMS functionality of the GGSN lies in providing IP connectivity for the IMS control (e.g. SIP, Diameter and H.248) and media.
The routers and switches in the IP infrastructure provide transport capabilities for both the control layer and connectivity layer traffic.

2.4.1 CSCF

The IP multimedia system is built around the Call Session Control Function (CSCF). There are three different CSCF types: the Interrogating CSCF (I-CSCF), the Proxy CSCF (P-CSCF) and the Serving CSCF (S-CSCF).

The P-CSCF is the first IMS point of contact for the User Equipment (UE). The P-CSCF forwards the SIP messages received from the UE to a CSCF in the home network (and vice versa). The P-CSCF may modify an outgoing request according to a set of provisioned rules defined by the network operator (e.g. address analysis and potential modification).

The I-CSCF function forms the entrance to the home network. It provides flexibility for selection of a S-CSCF, and may hide the inner topology of the home network from other networks.

The S-CSCF performs the session control services for the UE. This includes routing of originating sessions to external networks and routing of terminating sessions to IMS . access network (P-CSCF). The S-CSCF also decides whether a SIP Application Server is required to receive information related to an incoming SIP session request to ensure appropriate service handling. The decision at the S-CSCF is based on information received from the HSS (or other sources, e.g. application servers). The identifiers of the application server(s) are also received from the HSS.

All CSCF functions may generate charging information.

2.4.2 HSS/SLF

The Home Subscriber Server (HSS) is an evolution of the Home Location Register (HLR) and Authentication Center (AUC). It holds the subscriber's profile and keeps track which S-CSCF is handling the subscriber. It also supports functions like subscriber authentication and authorisation (AAA).

In networks with more than one HSS, the Subscriber Location Function (SLF) is used as well. The SLF assists the control layer with the location of the HSS server where subscription information resides for a specific subscriber.

2.4.3 MRF

The Media Resource Function (MRF) contains functionality to allow manipulation of multimedia streams. It can support functions like multi-party multimedia services, multimedia message playing and media conversion services. 3GPP has decided to split the MRF into a control part (MRFC) and a processing part (MRFP).

2.4.4 BGCF

The Breakout Gateway Control Function (BGCF) selects the network in which breakout to the GSTN network is to occur. That network may either be the home network, the visited or another network. If the interworking is performed in the home network, the BGCF selects a Media Gateway Control Function (MGCF). If the interworking is to be performed in another network, the BGCF may select another BGCF or an MGCF.

2.4.5 MGCF

The MGCF provides interworking between the SIP session control signalling from the IMS and ISUP/BICC call control signalling from the external GSTN networks. It also controls the media gateway that provides the actual user plane interworking (e.g. convert between AMR and PCM coded speech).

2.4.6 SG

The signalling gateway (SG) provides bearer interworking for the control signalling (e.g. ISUP/IP – ISUP/TDM).

2.4.7 MGW

The MGW provides media stream manipulation towards PLMN/PSTN/H.324M/H.323 networks.

2.4.8 PCF

The policy control function, which is co-located with a P-CSCF, terminates the Go interface. The Go interface between the GGSN and the PCF allows for bearer and session co-ordination (for the support of charging correlation); notification of radio bearer loss or modification; policing of the media flow destination and authorisation of the PDP contexts for that session.

2.4.9 IP version support

The Ericsson SIP based multimedia solution support IPv6 according to the 3GPP and IETF standards. In addition, the solution can be configured to support an IPv4 terminal (split terminal).

Note: In the IPv4 configuration, the Ericsson SIP based IP multimedia network behaves as IPv4 SIP proxies, the session and bearer levels are not correlated.

2.5 *Service Network*

2.6 *Service Network*

The Ericsson service network is built up from individual components that can co-exist with other service network nodes and concepts; or benefit can be gained from the Ericsson service network framework by utilising common subscriber management and common database handling.

2.6.1 Service Network Framework

SNF Service Network Framework is a framework for components residing in the service network. While an Ericsson concept, the SNF is an open, IP based application environment using standard protocols and industrial practice to deploy a Service Network.

Some of the concepts that the SNF is built around are:

Common Provisioning

There should be only one point of access for the end-user and one point of access for the operator to manipulate the applications data. The end-users to personalize his/her services, the operator provision the service and default data. The end-user should have a possibility to subscribe/unsubscribe to applications provided for the users by the operator. One of the key components for common provisioning is the common directory. The Common Directory provides a single point of access to subscriber data, resident in many different databases. Applications can now access information stored by many other applications in a secure manner.

Common Management

The Service Network Operational System (SNOS) is the single point of O&M for the Service Network. SNOS is a fault, performance and configuration management tool.

Enablers

Enablers aid in the development of Mobile Internet services, it provides multiple functions and features needed in the development and customisation of services. The enablers are vital in making a horizontally layered architecture possible. Each enabler interfaces with core and business support systems to provide APIs (Application Programmer Interface) for developers to use.

2.6.2 SIP Application Server

A key component of the SIP based IP multimedia service network is the SIP application server (SIP-AS). The Ericsson SIP-AS contains a service creation environment supporting software development kits (SDKs) which can be used to access the SIP applications, but also other enablers.
The SDK is built around Java interfaces supporting the J2EE application environment on the J2AS (Ericsson J2EE Application Server). The interfaces have a common look and feel making it easier for the developer to find functionality and classes in a well-structured environment.
The diagram below describes how the SDK provides integration to the different enabler components of the Service Network Framework.

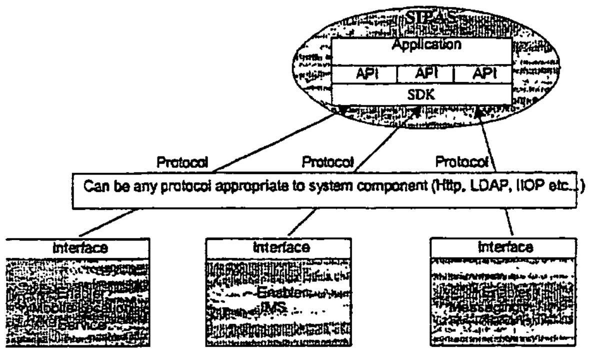


Figure 3 Example use of SIP-AS and enablers.

J2AS is implementing the J2EE standards and APIs to support the SIP protocol. Through the SDK, application developers do not need to know what middleware technology is implemented towards service enablers.

2.6.3 Presence and Instant Messaging

Ericsson's presence solution is based on presence information from a number of sources including CS domain, PS domain, mobile positioning centre, as well as information from the IMS. Instant messaging is communication between two users, or a community of users. Community based instant messaging is supported with the presence and instant messaging application.

Ericsson's system supports the presence and instant messaging (PIM) application on the J2EE SIP application server. The PIM application for IMS is integrated with Wireless Village (WV) PIM application in order to co-ordinate the mobile PIM communities and re-use of WV application infrastructure and databases. The PIM application supports the IETF "SIMPLE" approach for interworking with other PIM communities.

2.7 IP infrastructure

When providing an IP infrastructure for IMS, the objective is to let all services (e.g. IMS, fixed-line services, ISP services, content delivery transport) use the same IP infrastructure.

The basic assumption is that similar real-time characteristics are wanted for this multi service IP infrastructure as the characteristics that are offered with TDM and ATM networks.

The IP infrastructure is structured into two main tiers:

- Site tiers where network elements using local area technology are connected to the IP backbone.
- An IP backbone tier carries all traffic between the sites using wide area technology

2.7.1 Sites

A site has a local area IP infrastructure, which is connected to the various elements at the site and connects to the IP backbone trough edge routers. The site is typically using Fast Ethernet, Gigabit Ethernet, LAN switches etc, with capabilities of using VLAN techniques. The IP infrastructure of the site is duplicated to guarantee full availability.

The edge routers connect the site to the IP backbone, and contain advanced functions for defining a service agreement between the site IP network and the IP backbone. This includes e.g. MPLS LER function, 2547bis, BGP, and various filtering functions. A site typically uses a pair of edge routers, connected to different core routers in the IP backbone. The mapping of these different site types on the IP backbone network is shown in Figure 5.

Ericsson divides the sites into 3 site types capture the needs of an operator. Other site types can be defined depending of specific operator conditions.

- Primary Site
  is the most important site type and includes a "complete set" of functions needed for a UMTS network: IMS servers, media gateways, GPRS support nodes, and radio access network controllers. A primary site may also include a service network configuration.
  A network can have several primary sites for redundancy load distribution purposes.
- Secondary Site
  contains media gateways, GPRS support nodes, and radio access network controllers. If suitable in the network secondary sites can also have peering connections to other networks.
- Access Network Site
  includes media gateways and radio access controllers, and is used to concentrate the traffic before going to a primary or secondary site.

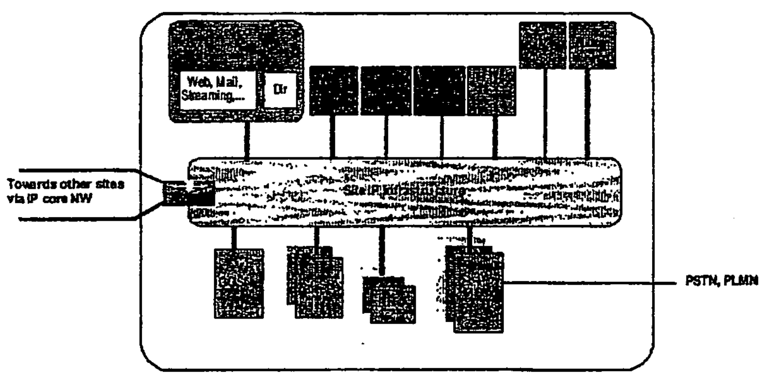


Figure 4 Example of a Primary Site

2.7.2 IP Backbone

The IP backbone is designed for simple high-speed packet transport, and is typically built with large backbone routers interconnected with fast links, e.g. Ericsson's AXI 520/580 routers interconnected with gigabit links.
The assumption here is that one single operator runs the IP backbone. The routers support different class of services and layer 3 VPNs.

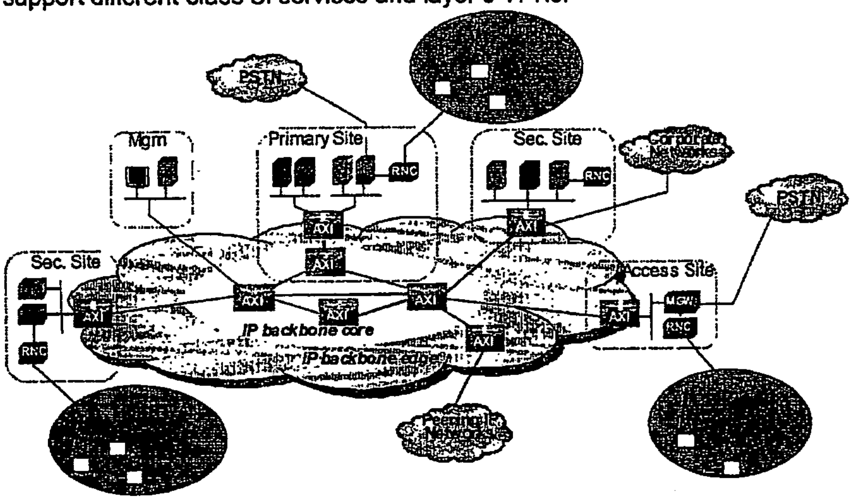


Figure 5 Sites interconnected by IP backbone

3 System concepts

3.1 IPv6

To provide enough IP addresses to all connected terminals it is envisaged that operators would provide IPv6 for end-user and applications. For the IP backbone there will be a long transition period when both IPv4 and IPv6 will be used. Ericsson products will include dual stack implementations. Features for handling IPv6 packets in different IPv4 tunnels will be provided.
3GPP IP Multimedia Subsystem has been defined to use IPv6 only. The diagram below shows on a high level where the impacts are in the existing system, mainly the terminals and GGSN. As peer-to-peer, machine to machine, push, presence, multimedia and mobile focused service communication become dominant in the market, use of IPv6 becomes a natural way forward.
The successful transition from a IPv4 dominant Internet to IPv6 will be a long process and proper and adequate transition mechanism will be a key enabler of such smooth migration. Ericsson has focused on a strategy where

- all IPv6 enabled nodes will be dual stack, including terminals,
- IPv6 will be introduced in the application layer first (e.g. IMS), thus enabling gradual and controlled introduction of IPv6 allowing gain in customer experience and knowledge on practical deployment of IPv6
- migration and transition mechanism are provided to co-exist with IPv4 deployed networks

- All intermediate steps are directed towards long term goal of moving into IPv6 in the transport plane, end to end and top to bottom when IPv6 is dominant in the market.

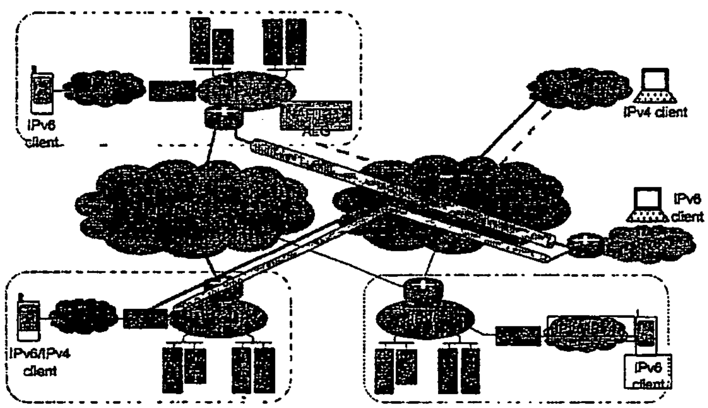


Figure 6 Possible rollout scenario

3.2 Quality of Service

The starting point for QoS should be the end-user and end-user satisfaction with provided services. This is crucial for real time IP services such as voice, video, and music. But to provide this satisfactory quality, a chain of mechanisms has to be engaged on the path from application to application. Ericsson's products provide a large number of advanced QoS features on the different segments of the chain.

3.2.1 QoS in UMTS

In this chain the radio and the resource control of radio resources is crucial. Ericsson's products for UTRAN and GPRS supports the advanced QoS handling required when volume IMS services are provided.

3GPP has standardised a QoS model and architecture, supported by a set of detailed attributes, to be used to configure and control the traffic from the mobile station to the network. Four traffic classes is defined:

- Conversational
- Streaming
- Interactive
- Background

Each class is defined to match a typical application scenario. Conversational traffic class is optimised to support a human-to-human real time conversation of voice or video, Streaming traffic class is optimised to support streaming video or audio applications, Interactive traffic class is optimised to support non real time, typically TCP based, interactive user flows. The purpose of background is to provide a data transport where there are no demands at all on delay.

Ericsson's UTRAN and GPRS products supports all traffic classes and provides efficient transport of the user flows over different transmission media, WCDMA, ATM, IP, without compromising the end-user quality.

3.2.2 QoS in IP backbone

Traditionally IP backbone networks had been implemented only relying on best effort forwarding and large enough bandwith. Together with the congestion avoidance mechanism in TCP, this is an adequate and cost effective principle if the data transported, or at least the greater part of the data, is non real time (TCP/IP) traffic.

3.2.2.1 Congestion control mechanisms

When IMS is implemented there will be a need for the IP backbone to efficiently transport large volumes of real time data. Note this real time data does not use TCP, and therefore will not be able to benefit from the congestion avoidance mechanisms in TCP.

Over provisioning of the IP backbone is therefore not considered a viable, general solution for providing QoS in an IP backbone supporting a volume of IMS services. Additional means for congestion control are necessary to improve network efficiency.

The extent to which over provisioning can be reduced depends on the grade of sophistication of the congestion control mechanisms. The amount of over provisioning needs to take into account fault situations (link or router failures), traffic concentrations in conjunction with 'abnormal' events, provision of capacity for best effort IP traffic (must never be completely starved), etc.

The congestion control mechanisms supported in Ericsson's IMS solution is:

- Admission control in the nodes processing IMS user data, i.e. GGSN, MGW, MRFP
- DiffServ differentiation in the edge and backbone routers
- Ability to define a bandwith limit when scheduling prioritised queues
- Ingress policing of external interfaces (where admission control is not trusted)
- Ability to use traffic engineering to control and dimension the network (e.g. by using MPLS)

3.2.2.2 Admission Control

The nodes involved in IMS user data transport over the IP backbone are GGSN, MGW and MRFP. To be able to dimension the IP backbone network and to avoid

overloading the backbone routers with real time traffic these nodes support an admission control function.

The data processing function in GGSN, MGW and MRFP have a flow context. For example GGSN identifies to which PDP context a packet belongs received over Gn: MGW, when gating between CS network and IP network, makes use of the circuit concept of ISUP/ISDN.

Before accepting a request for a new flow GGSN/MGW/MRFP check that there are available IP resources matching the required forwarding quality of the flow. The required forwarding quality corresponds to DiffServ PHB and is retrieved from UMTS QoS attributes (PDP context) or media type description (received from MGCF/MRFC). New flows are requested by the session/call control function: PDP context activation for GPRS and the bearer request that is as a part of IMS session establishment procedure (from MGCF and MRFC). If the resource utilisation limit for a specific forwarding quality is reached in GGSN, MGW or MRFP, requests for PDP context or bearer will be rejected.

When a request for a PDP context / bearer is accepted the required forwarding quality is stored in the context for the flow (GPRS PDP context/ bearer context). During user data transport the user flows will be 1) classified according to their required forwarding quality, fetched from the context 2) mapped to DiffServ PHB and 3) marked with corresponding DHCP.

The IP forwarding in edge router and core routers will only be aware of the PHB retrieved from DSCP of received IP packet. The flow context is transparent for the routers. The IP network should be dimensioned so that it has capacity to forward all DiffServ traffic of a certain class according to the defined PHB of the class as long as the configured resource utilisation limit is not exceeded in GGSN, MGW and MRFP.

3.2.2.3 DiffServ

DiffServ constitutes a corner stone for handling QoS at the IP-layer in the proposed QoS-solution. As said above the nodes involved in transport of high volume real time IP traffic perform the DiffServ marking of an IP-packet.

The edge and core routers are configured in such a way that the intended traffic scheduling and prioritisation is obtained for packets according to their DSCP. The IP packets are classified and forwarded to a suitable outbound queue, on which the scheduler operates.

To conform to the dimensioning rules defined for the network, the scheduler should be configured with a bandwidth value per queue, e.g. assigned a percentage of the total available bandwidth and possibly also a priority. It is essential that mechanisms in the router prevent starvation of low priority queues.

Ericsson's AXI 520/580 routers and the embedded router in MGW, MRFP and GGSN provide rich mechanisms for DiffServ based forwarding: e.g. queuing, scheduling and packet marking/dropping at the interface level. Ingress policing, queue selection,

precedence field rewrite/remark, Random Early Detection (RED) and strict priority queuing are other configurable functions.

3.2.2.4 Principles for network dimensioning, an example

There is a trade off between sophistication of the congestion control mechanisms and the extent to which over provisioning can be reduced. The cost to maintain and configure a sophisticated congestion control mechanism in a large network may be significant, but also the cost for transmission links if the amount of overprovisioning is high.

Ericsson has during the last years worked with the issues of dimensioning a multi service IP infrastructure, and found that the following principles is a good balance between complexity and IP resource efficiency:

The resource utilisation limit in a node is configured as allowed bandwidth per "forwarding quality" and interface. The interface is the interface in the edge router of the site that is used for the IP flow (see chapter 2.5.2). In case of VPN or redundant links, several interfaces may have to be configured in the admission control function. The accumulated load retrieved from UMTS QoS profile (GPRS) or the media type description (IMS) is compared with the configured bandwidth.

This model offloads the dimensioning tool and configuration management from configuring a full mesh network between all GGSNs, MGWs and MRFPs. The increase in level of overprovisioning is heavily dependant on number of sites and network topology, but simulations has shown that the increase will be a factor of 2 –3, compared to a full mesh network, for typical topologies.

3.2.2.4.1 Static versus dynamic network dimensioning

The dimensioning model and admission control principles described above assumes static configuration of the admission control functions of GGSN, MGW and MRFP as well as the interfaces of edge and core routers. The admission control function in a node will statically be assigned a fraction of the total bandwidth of an outbound interface of the edge router, the bandwidth level selected so that the interface is not overloaded and prevention of starvation of best effort traffic.

Ericsson recommends static resource configuration as the first step when building a multi service IP network, mainly because the lack of widespread multivendor resource control protocols and lack of full-scale experience of dynamic resource control algorithms for IP networks.

Dynamic handling may be introduced stepwise as the network grows and new resource control mechanisms evolve.

3.2.3 End-to-end QoS coordination of IMS sessions

The detailed mechanisms, segment by segment that allows the operator to control the quality provided to the end user of IMS service has been detailed above.

The last piece of the QoS puzzle is the feed back to the end-user. When establishing an IMS session to a remote IMS user, the end-user would like to know if there are

enough resources, all the way to remote application, before the session starts. The model for session establishment selected by 3GPP is based on QoS assured SIP protocol option. During session establishment SIP messages are exchanged informing the remote side if the bearer level QoS set-up was successful or not.

If for example an admission control function rejects the requested QoS level, SIP QoS assured protocol option supports that the side that fails informs the remote side of the failure. If this is not acceptable the IMS application in the terminal may terminate the session, but it is also possible for the two participants to try to continue on best effort basis, i.e. no longer using the QoS assured protocol option.

3.3 Security

3.3.1 Access Security

The IMS access security includes

- mutual authentication between the user and the IMS home network.
- confidentiality (optional) and integrity protection between the user and the IMS access network.

For the integrated terminal, the authentication is based on IMS AKA, following the 3GPP standards. For split terminals, which do not have access to the UICC, then standard SIP methods are supported, for example HTTP digest (HTTP basic is not recommended).

Note: The standardisation of the IMS access security is still in progress in the 3GPP.

3.3.2 Network Domain Security

The network domain security (NDS) provides Authentication; Replay protection; Integrity protection and Optionally confidentiality protection between network domains, and optionally within a network domain.

In order to support NDS between networks, Security Gateways (SEGs) are provisioned on the border of the operators network, protecting the trusted network from the outside, untrusted network.

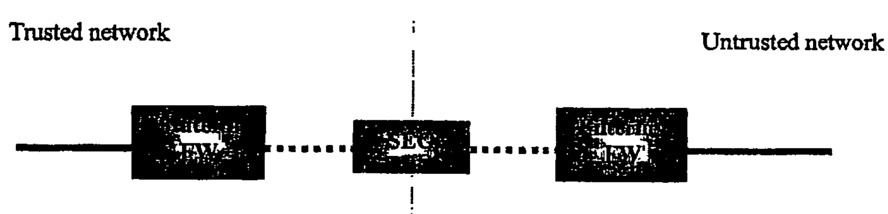


Figure 7 Network Domain Security Support

The security gateway functionality consists of

- a filtering firewall,
- the actual security gateway and
- a stateful firewall.

The task of the filtering firewall is to act as a first line defence against attacks from untrusted networks using basic packet filtering firewall capabilities. The filtering firewall can be e.g. an Edge Router.

IPsec VPN tunnels are used between co-operating SEGs. The stateful packet inspection firewall constitutes the second line defence for all the traffic not using IPsec VPN tunnelling, but also for the traffic coming via IPsec VPN tunnels as attacks through the tunnels are also possible. Stateful inspection of TCP sessions and UDP pseudo sessions is used to increase security.

Note: The standardisation of the Network Domain Security is still in progress in the 3GPP.

3.3.3 Site Security

A site must have the security functions that are necessary to support the operator's security policy. A site in this context is an IP infrastructure Primary Site described earlier including the IMS nodes.

The security functions below are useful functions to support a security policy.

3.3.3.1 Perimeter protection

Firewalls are used for perimeter protection and protect fully or partly against general IP threats.

Firewall filtering is done based on various parameters e.g. port numbers, IP addresses, interfaces (in, DMZ, out) and protocol (TCP, UDP, SCTP etc):

Ingress filtering should be used to check the source IP address to reduce the possibility of denial of service attacks. This should then be enabled in GGSN.

3.3.3.2 Hardening

Hardening refers to the process of modifying system resources to be more secure to protect internal nodes to some extent if the perimeter protection is penetrated or from insiders.

3.3.3.3 Audit trail recording

Audit trail recording or logging must be used in a system to be able to detect suspicious network activity.

Logging of relevant events defined by the security policy (e.g. access attempts, system changes, indication of network reconnaissance) is included in network nodes. Filtering is possible to apply in order to reduce the sizes of the log files. Those files shall be securely stored.

3.3.3.4 Intrusion Detection Systems

An IDS should be used to automate analysis of the audit trail files. Ideally a successful IDS can recognise both suspicious activity and denial-of-service activities and invoke countermeasures against them in real time.

3.4 Logical Networks

Different types of information are exchanged between the network elements. Traditionally, separate networks were used to handle these, e.g. STM/TDM, ATM, IP and SS7 networks, each with well-defined quality of service and with little or no connectivity among them, thus giving complete separation of traffic between the various information flows.
When these networks converge into a single multi service IP network in which IMS is a part, the IP infrastructure must be capable of handling all this information exchange. At the same time it must facilitate for necessary traffic separation and assurance of sufficient quality of service for the different traffic types. Ericsson has thoroughly studied these issues and has coined the term *logical networks* to conceptually divide the independent traffic types.
Each logical network encompasses a particular information flow between a designated set of functional entities residing in the network elements. Each logical network has both a set of requirements on the infrastructure with respect to security, connectivity, QoS, network availability etc and a set of characteristics that might influence other flows, such as bandwidth demand, burstiness and possible security threats to other logical networks.

<u>3.4.1 Mapping of logical networks to VPNs</u>

Ericssons definition of different logical networks is based on the different functions in the system (e.g. exchange of signalling information, O&M, exchange of user payload etc), their characteristics and requirements. Some logical networks have similar requirements and can be combined (e.g. SIP signalling and traditional SS7 signalling) other logical networks need to be separated (such as user payload from signalling and O&M traffic).

VPN technologies can be used to separately transport the different logical networks on the same physical infrastructure. Each logical network has different requirements on the IP network, so different VPN technologies should be used for different types of traffic. One simplified mapping is shown in the Figure 8. BGP/MPLS IETF RFC 2547bis (with extensions to IPv6) is used to separate logical networks into layer 3 VPNs in the backbone. Within the site, separation is performed with VLAN tagging. This gives similar separation characteristics as in an ATM network.
Also other mappings like e.g. IPSec VPNs can be used where privacy is required.In some cases VPNs can be avoided and substituted by filtering and access control lists in the routers and client nodes.

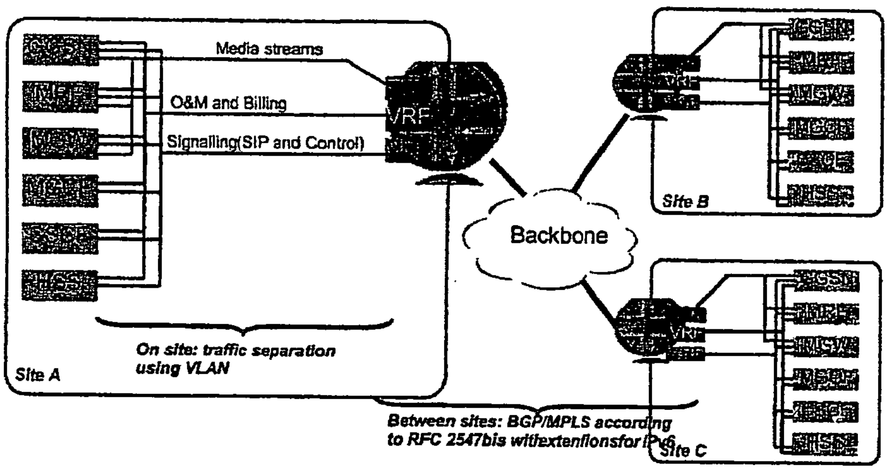


Figure 8 Logical Networks mapped to VPNs

Effective separation into VPNs must be supported in the network elements. The Ericsson platforms used in the IMS architecture are designed to support multiple connections to different networks and are thus ideally prepared for this task.

3.5 Addressing

The SIP based IP multimedia addressing is a collision of two familiar worlds. E.164 numbers, which is applicable in the 2G mobile networks, are available, and in addition, the "Universal Resource Locator" (URLs), which take the form of email addresses are supported. As an example, a valid SIP address might be sip:name.surname@company.com, which implies the use of DNS to map host and domain names to IP addresses. This is a key aspect of the integration of SIP with web and mail-enabled technologies, which are already familiar with the concepts of URIs and their interpretations.

The close connection between SIP and DNS facilitates interoperability with telephony systems as well as their addressing mechanisms. Support for ENUM (RFC 2916), which describes a means of resolving E.164 numbers to Internet resources (such as SIP URIs or IP addresses), allows SIP servers and clients to send and receive telephone numbers in place of SIP URIs in messages, and route them in a sensible fashion. Thus, examples of acceptable SIP URLs include:

sip:j.doe@big.com
sip:j.doe:secret@big.com;transport=tcp
sip:j.doe@big.com?subject=project
sip:+1-212-555-1212:1234@gateway.com;user=phone
sip:1212@gateway.com
sip:alice@10.1.2.3
sip:alice@registrar.com;method=REGISTER

Figure 9: Examples of valid SIP URLs

When an e.164 number is used, the IMS will attempt to translate the address to a SIP URL, for sensible routing. In the case that a SIP URL is not applicable, the IMS will forward the session establishment to the PSTN/ISDN network for continued routing.

3.6 Operation and Maintenance

Ericssons operation and maintenance solution includes element management and subnetwork management functionality. Standardised IRP interfaces supported.

3.6.1 Element Management HSS, CSCF (I-, S-, P-), MGCF, MRFC and, BGCF

Element Management functionality is provided from a Toolbox which can be launched from any standard web-browser. The following functionality is provided in the O&M Toolbox:

- Performance management. Dashboard application allows the user to select any available performance management counters for display in a GUI.
- Configuration Management. LDAP Browser for debugging, configuration management, system maintenance, and provisioning activities. Signalling stack GUI.
- Fault Management.
- Alarm and Notification web viewer.
- Security GUI is provided for user, password and permissions management
- Product inventory. Files containing hardware and software information that can be accessed from an external management system.

3.6.2 Element Management MGW, SG and, MRFP

The Element Management function is built on a thin client application using CORBA technology. It can run from a standard web browser on any computer, either locally or remotely. The Element Manager supports the following applications:

- Fault Management. Alarm log and alarm list are accessible from the element manager. Documented operational procedures are connected to each alarm, and are available on-line.
- Software Management. Software installation and upgrade can be performed manually via the element manager
- Equipment Management. Hardware product inventory information is provided in the element manager (XML file browsing)
- Configuration Management. GUI and scripting interfaces are available for configuring and controlling physical interfaces, link protocols and devices
- Performance Management. Performance Data IRP is provided to allow performance managers to control and read the counters. No support is provided in the element manager for control or reading of performance counters on the node level.

3.6.3 Subnetwork management

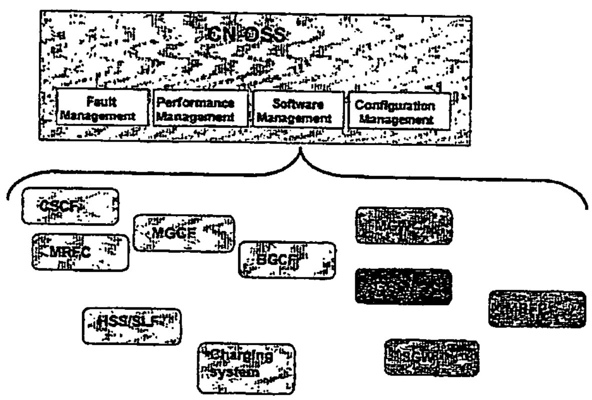
The CN-OSS subnetwork manager is available to manage the IP multimedia nodes, in addition to the tasks it performs in managing the traditional core network nodes already deployed.

Each of the applications in the CN-OSS solution are designed to:

- Provide centralised management for core network and IP Multimedia nodes

- Reduce time required to perform management tasks
- Reduce the time, and cost required, to build user competence
- Minimise the time that parts of the core network are not in service

The diagram below shows CN-OSS and the nodes that it manages.



The CN-OSS subnetwork manager provides the following functionality:

- Software management / upgrade.

- Fault Management, including:
    - Fault Management Mediation
    - Fault Management Presentation
    - Fault Manager Expert (FMX) (rule-based expert system)

- Performance Management.

- Configuration Management.

- Job Manager.
    - The Job Manager in CN-OSS provides the capability to define work activities that can be executed for nodes and node types managed by CN-OSS.

3.7 Transport of applications flows

When using the WCDMA technology to access the SIP based IP multimedia system, the SIP signalling and the media streams must be transported over the Universal Terrestrial Radio Access Network (UTRAN) and Packet Switched domain (PS domain). Efficient usage of air interface resource is enabled with careful choice of the

configuration used to transport the IMS bearers, including the SIP signalling and media.

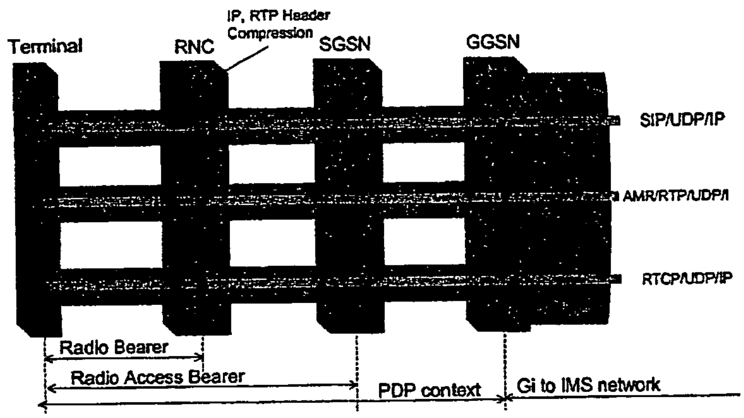


Figure 10 Transport of Application flows - Voice Only

The terminal chooses the means by which the Application flows should be transported over the UMTS network. This is dependant upon the application, the availability of radio bearers, charging considerations. Figure 10 shows a voice communication using the SIP based IP multimedia system. The SIP/SDP signalling, the AMR (RTP) media and the RTCP signalling has their own PDP contexts (and hence Radio access bearers). The PDP context for SIP/SDP signalling is always active to enable the terminal to receive IMS sessions, but during idle periods the RAB can be deactivated. The terminal selects the configuration of the access network required to supports the communication media's negotiated through the SIP signalling. When establishing the required configuration, the terminal provides the required Traffic flow template (TFT) information to the GGSN, instructing the GGSN to place the correct flows in the selected PDP contexts. The terminal must hence know how to map the signalling and media flow to PDP contexts.

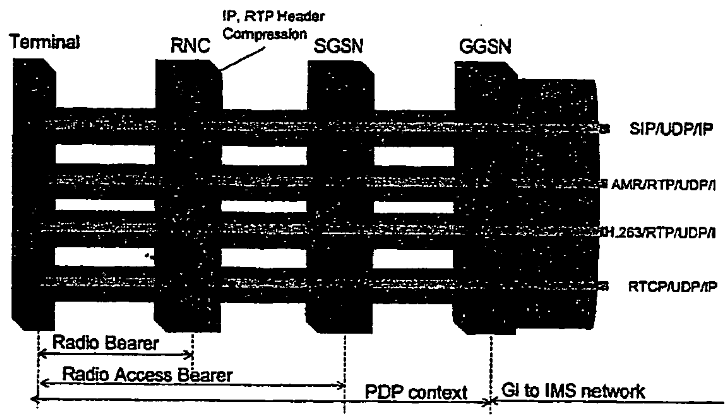


Figure 11 Transport of Application flows - Voice and Video

Figure 11 illustrates the scenario where video and voice communication occurs using the IMS. In this example application scenario, the Video component (H.263) is transported on a separate PDP context and RAB. The details for the support of this scenario require further investigation.

3.8 Charging

3.8.1 General

The way charging data is collected for the SIP based IP Multimedia services differ significantly from traditional telecom networks. The main reason behind this is that the signalling and the bearer normally do not follow the same path. Therefore, the basic principle is to handle charging for signalling and the bearer separately.

The following information can be input into the charging system:

- Transport information: Charging information of the used bearer resources is obtained from the PS domain.
- IP Multimedia session based information: Charging information provided by the SIP based IP multimedia system is collected from the SIP proxies. This includes also charging for SIP messages if they have added value (for instance, instant messages that can be sent in a SIP message)
- Content based information: This means the charging for the provisioning of information, like sports highlights, games. The charging information for content is delivered from the Service Network.

The following charging models can be supported:

- The calling party may incur charges entirely for both the IMS level charging and the transport (GPRS) level charging at calling and called party sides. Initially,

operators may require inter-operator agreements to enable this until the standards are stable.

- The calling party incurs transport level charging on the calling party's side only and the entire charges related to the IMS session level. In this charging model, the called party incurs the transport level charging associated with that session of the called party's side. The called party pays for its own transport level resources.

Each media component within an IMS session can be charged separately. If the called party request additional media components with regards to the initial request from the calling party, then the called party can be charged for these additional components.

In the case of roaming, the called party incurs charges up to the home network of the called party at the transport level. The latter incurs additional charges due to roaming from the home network to the visited network.

The transport and the session charging will occur independently if no binding mechanism exists to correlate the charging information. The charging correlation is under investigation in 3GPP but a standardised solution has not yet been defined. It will be supported when defined. When support terminals using IPv4 to access the SIP based multimedia network, the binding will not occur.

3.8.2 Prepaid/Postpaid Convergence

The dominating charging trend today is the rapid growth in PrePaid subscribers. The next foreseeable step is that all network services will be rated in real-time, which would mean a convergence of PrePaid and classic PostPaid. It would also mean that real-time rating becomes a core function in the network rather than a service. The Pre-Paid/PostPaid convergence allows the same charging, administration and rating systems to be used for all PSTN, GSM, UMTS and IP subscribers.

3.8.3 Online Charging

Online charging is commonly used today for GSM prepaid subscribers. This is also expected to be the case for prepaid IP Multimedia users. Additionally, with the convergence of prepaid and postpaid, all users will have the ability to have online cost control. The exact mechanism for the online charging is still being standardised.

3.8.4 Offline Charging

Charging information can be collected in different levels offline, producing CDRs that can be transported to the charging system.

3.9 Interworking

- Input from LMF (Janne Soutola) on Friday

4 Product view

4.1 Ericsson products

Considering the complex functional IMS architecture, it is essential to do a smart mapping between logical functions and actual products. A scalable and flexible IMS solution should be used.

It is also reasonable to believe that the first commercial deployments of IMS products will start on a smaller scale. It is therefore essential to have cost-efficient entry-level solutions that will still provide enough capacity to handle the initial IMS traffic. As the market evolves, more flexible and powerful configurations will be required.

The Ericsson IMS product line will fulfil both the scalability and the flexibility requirements by implementing the IMS functions in an integrated, yet modular, way by using the following guidelines:

- provide scalability from entry-level (integrated) to high-end (distributed) configurations
- flexible software architecture (e.g. separate S-CSCF, I-CSCF and P-CSCF modules) allowing both combined and distributed solutions, where and when needed
- implement servers on TSP
- implement gateway and media functions on CPP for media and signalling gateways and media mixing

4.1.1 TSP

TSP is a platform for server and control network elements and is deployed in TDMA and CDMA networks. TSP is a suitable platform for nodes that require high availability and scalability such as the IMS nodes.

The architecture of TSP consists of functional units embedded in a framework of open interfaces. This allows the same software to be run on different hardware. The TSP uses primarily commercial components. Ericsson develops the most essential components for adding value in terms of robustness and scalability.

Linux is supported in the TSP processor cluster. Functions for scalability and availability are then added.

4.1.2 CPP

CPP is a generic platform for small to medium scale telecom applications. It has a robust distributed real-time telecom control system and low-cost ATM, TDM and IP transport.

CPP is used today in Ericsson products (e.g. RNC, BTS, MGW) in the first 3G systems rolled out around the world.

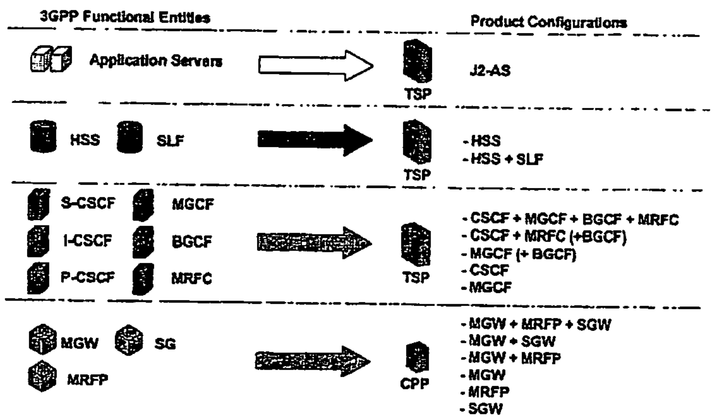


Figure 12 Mapping of IMS functions to products

Ericsson IP-works develops DNS and DHCP products adapted to mobile networks.

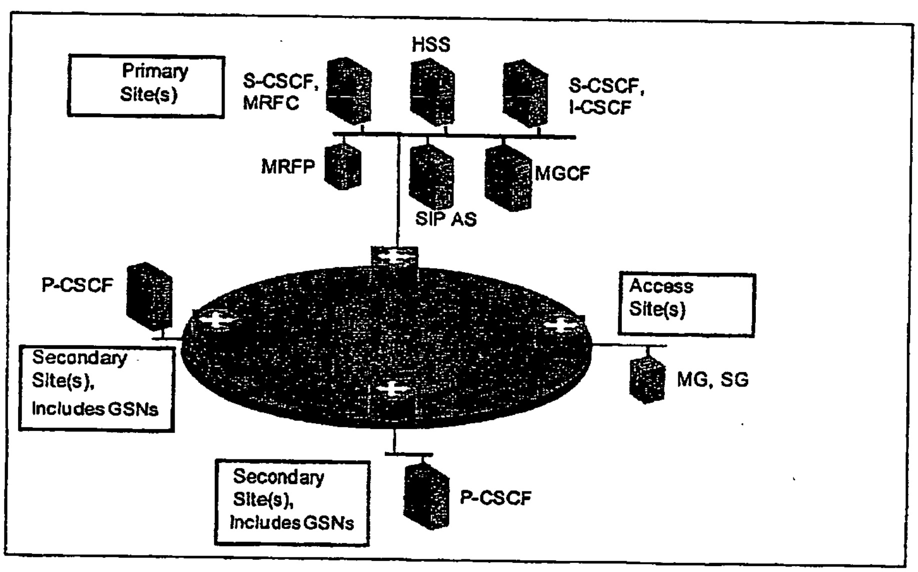


Figure 13 IMS commercial system configuration

Figure 13 shows an example of a configuration able to handle higher volumes of IMS traffic. We can distinguish primary, secondary and access sites. The secondary sites

could host P-CSCF functionality to handle the interface to the UE and forward the sessions to the appropriate primary sites. The primary sites host the functions providing the actual session control and services. The access site handles the interface to external GSTNs.

Figure 13 obviously shows only one possible scenario. Using the flexibility of our IMS product offering, many other configurations can be created, suited to specific network deployment scenarios.

4.2 Ericsson Joint Ventures

The GGSN J20 is a carrier-grade and router-based GGSN. It is a joint venture between Ericsson and Juniper Networks. This GGSN functionality is built on the platform of the Juniper Networks M20 Internet backbone router.

4.3 Ericsson partner products

For the IP infrastructure in the site Ericsson use products from our partners Netscreen for Firewalls, and Extreme for LAN switches. These products are well suited for carrier class networks with high availability architecture and hardware supported filtering and forwarding. The partnership includes identifying and implementing specific mobile network requirements.

The partnership with Juniper Networks mentioned earlier for GGSN aims also to provide carrier class router family with enhancement for the mobile core network. Edge routers are normally based on the AXI520 series (equal to Juniper M20, M40 series). A new GGSN product J20 has been developed based on the Juniper router platform. For the core routers in the backbone either the AXI520 or AXI580 (equal to Juniper M20, M40, M160) can be used. The partnership between Ericsson and Juniper combines unique competence about mobile systems with unique competence about how to build carrier class routers.

Products from other vendors might be used as well if they support the same functions as the products from the partners.

5 Migration

5.1 Overview

The path to supporting a conversational SIP based IP multimedia system are:

- Demonstration System
  The SIP based IP multimedia system demonstrating the SIP capabilities and is available on the Ericsson premisis.
- SIP based IP multimedia Trial System
  Available on the operators premises, and is available for IPv4.
- Non conversational SIP based IP multimedia subsystem
  A commercial system support best effort multimedia over the PS domain and voice over the CS domain.

- Conversational SIP based IP multimedia subsystem
  A commercial system supporting conversational multimedia over the PS domain.

5.2 *Trial system*

The trial site will provide full IMS functionality, including support for e.g. video-conferencing and interworking with GSTNs. It will also provide a local service execution environment on the CSCF. Figure 15 below shows an example IMS product configuration for trial systems:

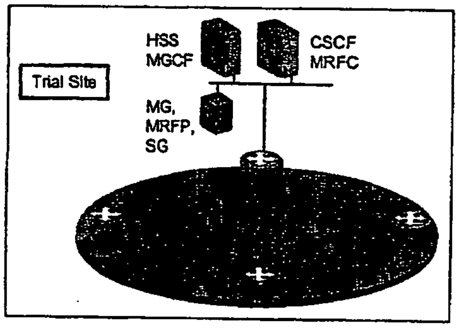


Figure 14 IMS trial system configuration

5.3 *Non-conversational SIP based IP Multimedia system*

A step towards a real-time conversational SIP based IP multimedia system is shown in Figure 15. In this migration step, the IMS part of the network is an IMS compliant subset of the network. As the terminals will be dual mode, supporting the real-time voice to be transported over the CS domain.

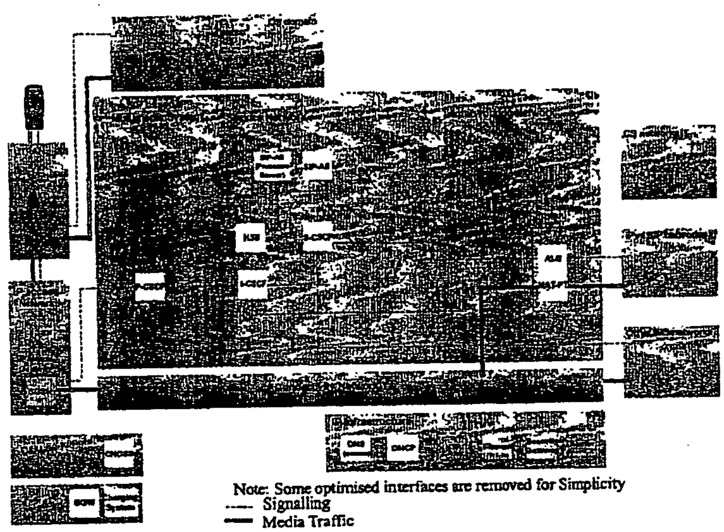


Figure 15 Migration architecture

The main features of the proposed migration network are:

- Compliant to IMS. A subset which allows for the addition of the PCF, MRFC/MRFP and interworking when increasing the capabilities of the IMS network
- Support for presence and instant messaging
- Support for services supported with the bearers which can be created with the deployed radio access network
- Simplified charging model: Each party pays their own transport usage, without correlation between session and transport.

PTO/SB/16 (6-95)
Approved for use through 04/11/98. OMB 0651-0037
Patent and Trademark Office, U.S. DEPARTMENT OF COMMERCE

PROVISIONAL APPLICATION COVER SHEET

This is a request for filing a PROVISIONAL APPLICATION under 37 CFR 1.53 (c).

Docket Number	2380-611	Type a plus sign (+) inside this box→	+

JC996 U.S. PTO

INVENTOR(S)/APPLICANT(S)

LAST NAME	FIRST NAME	MIDDLE INITIAL	RESIDENCE (CITY AND EITHER STATE OR FOREIGN COUNTRY)
BERGENLID	Lars	Herbert	Sollentuna, Sweden
OLSSON	Magnus		Spånga, Sweden

JC979 U.S. PTO 60/354483 02/08/02

TITLE OF THE INVENTION (280 characters)

IMS VOICE SERVICES

CORRESPONDENCE ADDRESS

John R. Lastova
NIXON & VANDERHYE P.C.
1100 North Glebe Road
8th Floor
Arlington

STATE	Virginia	ZIP CODE	22201	COUNTRY	U.S.A.

ENCLOSED APPLICATION PARTS *(check all that apply)*

☒ Specification — Number of Pages: 44
☐ Drawing(s) — Number of Sheets:
☐ Applicant claims "small entity" status.
☐ "Small entity" statement attached.
☐ Other (specify)

METHOD OF PAYMENT *(check one)*

☒ A check or money order is enclosed to cover the Provisional filing fees ($160.00)/($80.00)

☐ The commissioner is hereby authorized to charge filing fees and credit Deposit Account Number 14-1140

PROVISIONAL FILING FEE AMOUNT ($): 160.00

The invention was made by an agency of the United States Government or under a contact with an agency of the United States Government.

☒ No.

☐ Yes, the name of the U.S. Government agency and the Government contract number are:

Respectfully submitted,
SIGNATURE

DATE: February 8, 2002

REGISTRATION NO. *(if appropriate)*: 33,149

TYPED or PRINTED NAME: John R. Lastova

☐ Additional inventors are being named on separately numbered sheets attached hereto.

PROVISIONAL APPLICATION FILING ONLY

Burden Hour Statement: This form is estimated to take .2 hours to complete. Time will vary depending upon the needs of the individual case. Any comments on the amount of time you are required to complete this form should be sent to the Office of Assistance Quality and Enhancement Division. Patent and Trademark Office, Washington, DC 20231, and to the Office of Information and Regulatory Affairs, Office of Management and Budget (Project 0651-0037), Washington, DC 20503. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Assistant Commissioner for Patents, Washington, DC 20231.

601159